Predicting tryptic cleavage from proteomics data using decision tree ensembles

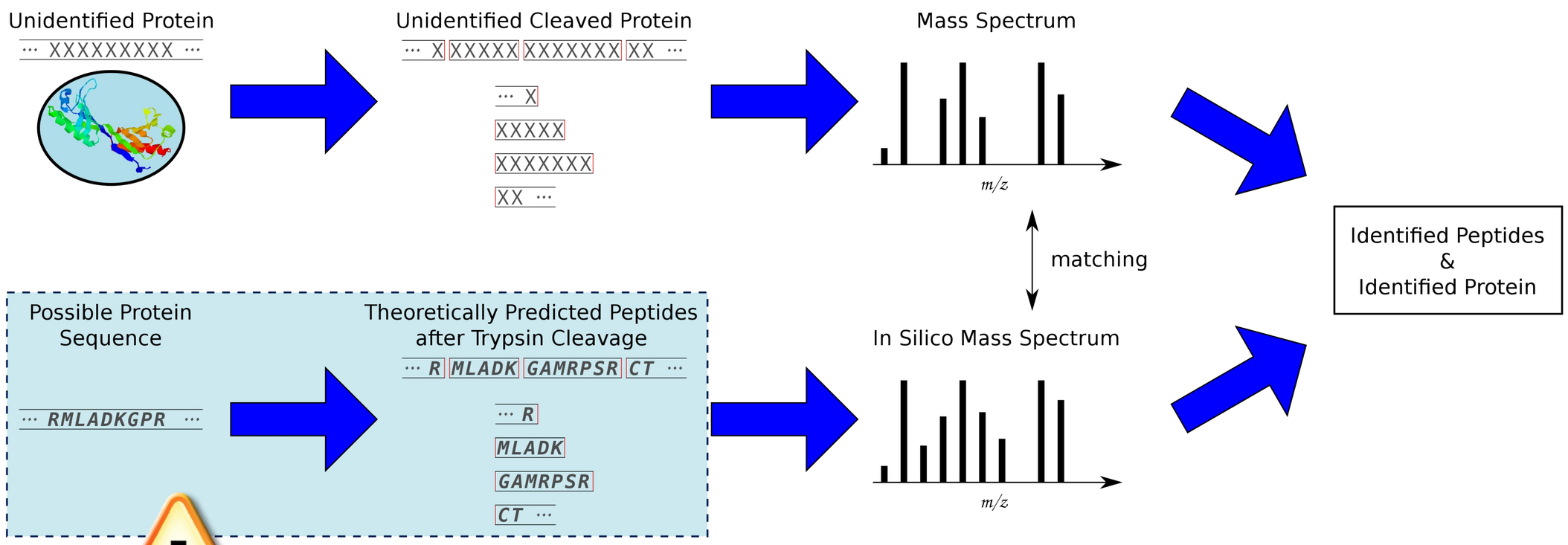
Thomas Fannes, Elien Vandermarliere, Leander Schietgat,
Sven Degroeve, Kurt De Grave, Lennart Martens, Jan Ramon

Journal of Proteome Research **12** (5) 2253-2259 (2013)

Contact: thomas.fannes@cs.kuleuven.be

Protein Identification

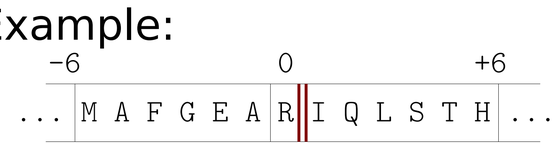
Goal: Facilitate identifying unknown proteins by trypsin cleavage and mass spectrometry



Accurately predict whether trypsin will cut a given sequence at a given position

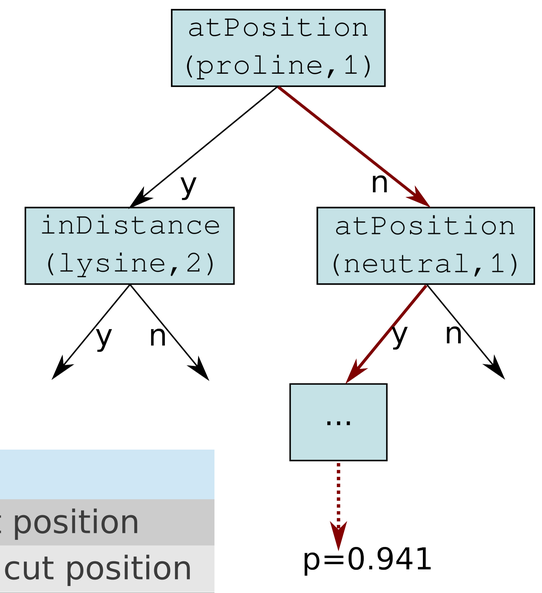
CP-DT: Cleavage Prediction with Decision Trees

Example:



p = probability of cleavage after R or K in this sequence window

Decision tree based on tests on properties of the amino acids in the window.



Amino Acid Properties
Name: alanine, arginine, ...
Size: tiny, small, medium, large
Charge: neutral, basic-positive, acidic-negative
Polarity: polar, non-polar
Ring: no, aromatic, imidazole
Atom-type: hydrocarbon, thiol, carboxyl, hydrogen, amine, amide

Tests
inDistance(p,d): Property p within distance d of cut position
atPosition(p,d): Property p at position p relative to cut position
isSpecies(s): species s of protein (Human, Mouse, Yeast. E.Coli)

We use a random forest, an ensemble of 100 decision trees.

Datasets

Dataset		Size	Species	Ratio +/-
Training	Pride	681193	Human, Yeast, Mouse	7.173
Test	iPRG	9694	E. Coli	5.105
	CPTAC	23842	Yeast	10.043
	MS_LIMS	26079	Human	20.769

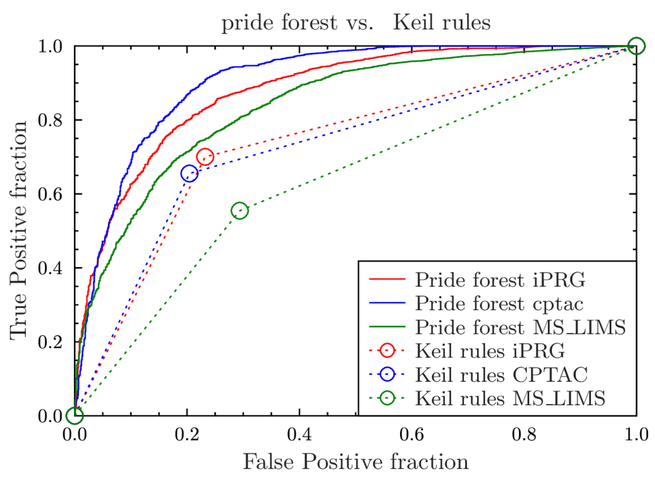
State-of-the-art Keil Rules

The sequence will not be cleaved if the window around the position has a similar structure as defined in the following rules:

		Relative position:				
-3	-2	-1	0	1	2	3
			[RK]	[P]		
			[RK]	[RK]		
		[DE]	[RK]			
			[RK]	[DE]		
			[RK]		[DE]	[DE]
[DE]	[DE]		[RK]			
	[DE]		[RK]		[DE]	

Keil, B (1992) Specificity of Proteolysis, *Springer*

Experimental Results



Inter-dataset results: (AUROC %)

Trained on \ Tested on	CPTAC	iPRG	PRIDE Mouse	PRIDE Human	PRIDE Yeast	PRIDE All	MS_LIMS
CPTAC	90.6	88.1	87.3	65.6	83.1	70.3	82.7
iPRG	86.5	88.5	85.9	67.9	80.9	70.9	76.7
PRIDE Mouse	89.2	89.6	96.2	69.3	83.4	73.5	80.5
PRIDE Human	85.9	81.3	85.1	97.0	83.8	95.5	80.9
PRIDE Yeast	89.4	83.4	86.4	72.5	98.4	78.7	82.8
PRIDE All	89.6	84.6	98.3	97.8	98.8	97.2	82.0
MS_LIMS	82.7	77.2	79.9	62.9	80.0	67.2	92.6

Interactive Webtool

Free CP-DT webtool at:
http://dtai.cs.kuleuven.be/trypsin

DTAI Declaratieve Talen en Artificiële Intelligentie
VIB
InSPECtor
iwt
erc Starting Grant 240186 MiGraNT

KU Leuven
Department of Computer Science